PSYCHOLOGY AND EDUCATION (2021) ISSN: 0033-3077 Volume: 58(3): Pages: 670-673
Article Received: 13th September, 2020; Article Revised: 25th January, 2021; Article Accepted: 12th February, 2021

Using Data Mining with C45 Algorithm for Student Data Classification

Iwan Rijayana*, Muhammad Ikhsanul Fikri, Ilham Fachrur Razaq, Rikza Syahbana Achlafass, Rizki Noorreza Allshal
Faculty of Engineering, Widyatama University, Indonesia
*iwan.rijayana@widyatama.ac.id

ABSTRACT

Analyzing data usually takes a long time, and is very complicated in its processing, even though the results are needed for decision making. For this reason, data mining is needed that can process data analysis using the software so that it finds a pattern and rules in the data set. Data mining can analyze data into information in the form of patterns that have meaning for decision support. One technique in data mining is data classification using a decision tree with the C45 Algorithm method. In this research, data mining software is created using the C45 algorithm method. The result is to be able to find a model or function that will explain or differentiate data classes, intending to be able to estimate the class of an object whose label is unknown.

Keywords
Data mining, classification, C45 algorithm, decision tree, data analysis

Introduction

Data mining is an automatic analysis of large or complex amounts of data to find important patterns or trends that are usually not recognized [1]-[2]. The results of the data mining application are evaluated to find new information/knowledge that is interesting and valuable to the company and then visualized to make it easier for users to choose the information that has meaning for decision support.

One of the processes in data mining is classification, the classification is given several records called training sets, which consist of several attributes, one attribute shows the class for the record. The purpose of classification is to find a model of the training set that distinguishes records into the appropriate category or class. One method used in classification is classification using a decision tree, with this method the training set is recursively divided until all parts contain all records from the same class. One commonly used decision tree algorithm is C45.

Data mining classification is used to look for patterns by analyzing a set of data sets that describe and differentiate data classes. This final project will analyze the decision tree method as one of the methodologies in data mining classification using the C45 algorithm, as well as the difficulty in seeing some of the scores of student courses affecting the cumulative achievement index of the student.

Methodology

Data Mining Method [3]-[4]

There are several methods or functions of data mining that can be used to find, explore, and knowledge. There are 4 main functions in data mining, including:

1. Predictive Modeling, this method is to build a model to predict a value that has certain characteristics. The predictive modeling method can be further grouped into two subcategories, namely: classification and regression. Classification is used to predict the value of discrete variables (such as predicting online users who will buy on a web site). Unlike the regression classification used to predict the value of a continuous variable (such as predicting future stock prices).
2. Association Analysis, also called market basket analysis where this function identifies product items that are likely to be purchased by consumers along with other products.
3. Clustering this method is to group homogeneous/similar data so that data in the same cluster have a lot in common compared to data in different clusters. Examples of clustering such as grouping documents based on the topic.

Article Received: 13th September, 2020; Article Revised: 25th January, 2021; Article Accepted: 12th February, 2021

4. Anomaly Detection, this method is to find anomalies or outliers, which are very different data from other data. An example of anomaly detection is determining an attack on a computer network.

Classification [5]

The classification algorithm seeks to find models that have high accuracy or low error rates when the model is applied to the test set. Some applications of the classification include direct marketing, detecting credit card fraud, medical diagnosis, and others.

The data classification process can be divided into 2 stages, namely:

a. Model Learning / Development
   Each record in the training data is analyzed based on the attribute values, using a classification algorithm to get the model.
b. Classification
   At this stage, the test data is used to determine the accuracy of the resulting model. If the level of accuracy obtained matches the specified value, the model can be used to classify new data records that have not been trained or tested before.

To improve the accuracy and efficiency of the classification process, there are several steps in processing the data, namely:

a. Data Cleaning

Data cleaning is the processing of data to eliminate noise and handling missing values in a record.

b. Relevance Analysis

At this stage, redundant or less related attributes are removed from the classification process to be performed. Relevance analysis can improve classification efficiency because the time needed for learning is less than the learning process for data with attributes that are still complete (redundancy still exists).

c. Data Transformation

The data can be generalized to data at a higher level. For example, by doing discretization of attributes with continuous values. Learning of generalized data can reduce the complexity of learning that must be done because the size of the data that must be processed is smaller.

d. Artificial Neural Network

The artificial neural network is a collection of processing nodes or units that map input-output, where the nodes are connected by links that have weights. During the training process, the network conducts learning by adjusting the weight value, so that it can predict the value of a class correctly.

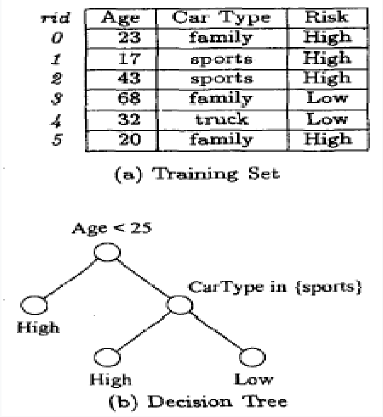
The weakness of ANN is the difficulty of processes in ANN to be understood. We cannot know for certain the meaning behind the weight values. Nevertheless, ANN has a relatively high level of accuracy for data that has never been inputted before, when compared to other classification methods. Also, ANN has a tolerance for data that contains errors (noise).

e. Fuzzy Logic Approach

Fuzzy logic is a system that allows boundaries that are not rigid (blurred) to define data. In the fuzzy set, there is a membership value between 0 to 1 which indicates the size of an attribute. The weakness of fuzzy logic is the cutting of attributes with continuous values. While its strength is its ability to classify with a high level of abstraction.

Decision Tree [6]-[7]

A decision tree is a flowchart that looks like a tree structure, as shown in Figure 1 (b). The decision tree recursively divides training data so that each part consists of the same class data. At each non-leaf node, there is a split point, which is a test of one or more attributes that show how the data is shared. Figure 1 (b) shows how the decision tree classifier based on the training specified in Figure 1 (a), (age <25) and (car type ε {sport}) are two split points that divide the notes into High and Low data risks. Decision trees can divide data that has not been classified into High risk or Low-Risk classes.



rid	Age	Car Type	Risk
0	23	family	High
1	17	sports	High
2	43	sports	High
3	68	family	Low
4	32	truck	Low
5	20	family	High

(a) Training Set

(b) Decision Tree

Figure 1. Decision tree

Article Received: 13th September, 2020; Article Revised: 25th January, 2021; Article Accepted: 12th February, 2021

Test Case

From the existing Student data, here only take several course values as a comparison to one's GPA. The GPA value is a class attribute; the GPA class is divided into 4 namely:
1. A: Very good (≥ 3)
2. B: Good (≥ 2,5)
3. C: Less (≥ 2)
4. D: Very Poor (<2)

For courses, the authors only take 9 types of courses as attributes, namely:
1. Management Fundamental (Semester I)
2. Accountant I (Semester I)
3. Accountant II (Semester II)
4. Business Fundamental (Semester II)
5. Financial Management (Semester III)
6. Human Resource Development (Semester III)
7. Business Planning (Semester III)
8. Research Operational (Semester IV)
9. Decision Support Theory (Semester IV)

The above course is a subject that must be taken by every student each semester. The author takes courses only up to semester IV and an average of 2 courses, so lecturers can see the GPA and determine what courses should be repeated or taken so that a student's GPA can increase. The courses that the authors take are subjects that are interconnected with one another or in other words as a prerequisite course, for example, to be able to take the Basic Mathematics II course a student must first graduate in the Basic Mathematics I course.

For more details, we see the following flow:
1. Accountant I → Accountant II → Financial Management → Research Operation
2. Management Fundamental → Business Fundamental → Business Planning → Decision Support Theory
3. Accountant I →. Human Resource Development

Results and Discussion

Table 1. Test results

No. Test	Courses	Training Error Rate	Time (sec)	Testing Error Rate	Time (sec)
1	Accountant I Accountant II Financial Management Research Operation	19,75 %	37	26,00 %	2
2	Management Fundamental Accountant I Accountant II Business Fundamental Financial Management Human Resource Development Business Planning Research Operational Decision Support Theory	6,00 %	73	47,00 %	6
3	Management Fundamental Business Fundamental Business Planning Decision Support Theory	22,15 %	17	49,00 %	1
4	Accountant I Accountant II Management Fundamental Business Fundamental	16,93 %	36	37,00 %	2
5	Accountant I Accountant II	16,00 %	27	45,00 %	2

Article Received: 13th September, 2020; Article Revised: 25th January, 2021; Article Accepted: 12th February, 2021

No. Test	Courses	Training		Testing	
		Error Rate	Time (sec)	Error Rate	Time (sec)
6	Financial Management Research Operation Decision Support Theory Management Fundamental Accountant I Business Fundamental Financial Management Research Operation	14,75 %	29	47,00 %	1

From the test results above, we can see the percentage of several courses on a student's GPA. The percentage of error rates generated on the results of testing is on average under 50%, some even 26%. That indicates that the resulting rule is good. The results were obtained from the training data on the subjects of Accountant I, Accountant II, Financial Management, and Research Operation. From the training data in the course, it turns out that it produces rules that are used for testing data with a very small percentage of error rate. The greater the percentage of error rate values generated in the testing data, the resulting rule is not good. The smaller percentage of error rate generated in the testing data, the better rule.

Conclusion

1. Data mining can be implemented by the decision tree method using the C4.5 algorithm.
2. Determination of training data will determine the level of accuracy of the tree made.
3. The time of tree construction is directly proportional to the amount of training data and the large number of attributes used at the time of tree development.
4. The percentage of tree truth is strongly influenced by the training data used to build the tree model.
5. Guardian lecturers can find out what courses can affect a student's GPA.

References

[1] Mehmed Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms”, Wiley-IEEE Press, 2020.

[2] Rahman, N. (2018). Data Mining Techniques and Applications: A Ten-Year Update. International Journal of Strategic Information Technology and Applications, 9(1), 78-97.

[3] Bergmeir, P. (2018). Enhanced machine learning and data mining methods for Analysing large hybrid electric vehicle fleets based on load spectrum data. Springer Fachmedien Wiesbaden.

[4] Karna, H., Vicković, L., & Gotovac, S. (2019). Application of data mining methods for effort estimation of software projects. Software: Practice and Experience, 49(2), 171-191.

[5] Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, 500-508.

[6] Islam, M., & Habib, M. (2015). A data mining approach to predict prospective business sectors for lending in retail banking using decision tree. International Journal of Data Mining & Knowledge Management Process, 5(2), 13-22.

[1] Kretowski, M. (2019). Evolutionary decision trees in large-scale data mining. Springer International Publishing.